Self-Supervised Convolutional Neural Networks for Fast Gesture Recognition in Human-Robot Interaction

1st Fotini Patrona
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
fotinip@aiia.csd.auth.gr

2nd Ioannis Mademlis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
imademlis@csd.auth.gr

3rd Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
pitas@csd.auth.gr

***Abstract*—Current autonomous systems (e.g., self-driving cars, autonomous drones, consumer robots, etc.) can already perform a wide variety of tasks and are predicted to be able to collaboratively assist humans in the near future. Thus, the need for efficient human-robot interaction (HRI) methods is greatly increasing. Gesture recognition is an effective HRI approach, since many robots are equipped with cameras and computer vision algorithms have progressed significantly in recent years, with advanced Deep Neural Networks (DNNs) being able to be executed on-board an autonomous system. However, computational/memory limitations are still significant for embedded AI methods, rendering the increase of DNN accuracy without imposing a penalty on runtime requirements a very important research priority. This paper investigates self-supervised DNN pretraining for a novel pretext task, relying on spatiotemporal video frame compression via tensor decomposition and low-rank approximation, as a means to augment gesture recognition performance, without inducing any runtime overhead during the inference stage. Thus, the method permits the use of less complex and much faster neural architectures that are well-suited to robotics applications and HRI. Quantitative evaluation on a gesture recognition dataset for autonomous Unmanned Aerial Vehicle (UAV) handling demonstrates the effectiveness and real-time performance of the proposed method on embedded AI compute hardware.**

***Index Terms*—Self-Supervised Learning, Human-Robot Interaction, Gesture Recognition, Convolutional Neural Networks, Tensor Decomposition, Low-Rank Approximation, Autonomous Unmanned Aerial Vehicles**

I. Introduction

Human-Robot Interaction (HRI) is a timely area of research, especially since the beginning of the current commercialization wave of autonomous systems for various application domains (e.g., self-driving cars, autonomous drones/UAVs, consumer robots, etc.) [14]–[17], [24], [25], [30]. In this context of increasing and ubiquitous autonomy, HRI methods permit robots and humans to collaborate in order to perform a task faster and more precisely than the human alone is able to.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No871479 (AERIAL-CORE).

Since most autonomous systems are equipped with a camera, gesture recognition is a an effective means to HRI. Given a sequence of video frames captured from an RGB camera, gesture recognition methods aim to predict a gesture class belonging to a predefined set of gestures, thus classifying an entire video sequence. However, this is not always an easy task, as the human performing gestures may appear in different working scenes and under varying scale, clothing and lighting conditions, which significantly affect the performance of gesture recognition methods. Complex Deep Neural Networks (DNNs) have proven able to partially overcome such difficulties in recent years, at a rather high computational cost. Several different neural architectures have been proposed for gesture recognition over time, such as combining Convolutional Neural Networks with Long Short-Term Memory networks (CNN-LSTM), 3D CNNs, CNN-LSTMs that process precomputed 2D human body skeletons [26] instead of the raw RGB video frames [27], etc. All of these alternatives, however, face execution difficulties when real-time operation on embedded AI platforms is required, as is typically the case with autonomous systems, due to unavoidable computational/memory hardware limitations.

This paper explores a special form of unsupervised learning, called *self-supervised learning* (SSL), as a possible answer to these difficulties. SSL focuses on extracting high-level, semantic visual representations from the input data by leveraging automatically created pseudo-labels. This is performed in a DNN pretraining stage using a so-called *pretext task*, i.e., learning to map variants of the training input data to pseudo-labels that are being automatically generated from the data themselves. Pretraining the DNN in a regular supervised manner on a suitable pretext task enforces the network to learn improved context-invariant features, easily transferable to another desired *downstream task*, such as gesture recognition, thus augmenting its performance on the latter one by reducing overfitting. In essence, SSL by pretext pretraining provides us with a better DNN parameter initialization to be used when training for the downstream task, therefore giving rise to increased accuracy at the inference stage, without *any*

architectural modification and/or runtime overhead. Evidently, this is an approach well-suited to embedded AI applications, where computational resources are limited when the trained DNN is deployed, but are potentially plentiful before (at the training stage).

The most prominent types of pretext tasks are the generative and the discriminative ones: the first ones involve content generation (e.g., GANs [10], colorization [42]), while the latter ones focus on context structure (e.g., jigsaw puzzles [1] and geometric transformations [6]) or context similarity [39]. In general, pretext tasks where the DNN is pretrained on images or videos are usually exploited for downstream tasks like object detection [21], image classification [43], or image segmentation [6], while pretext pretraining on videos is mostly used for action/gesture recognition from videos [36].

In this paper, a novel pretext task is proposed that leverages temporal video information by embedding it in each spatial 2D video frame representation, using tensor decomposition. The truncated HOSVD [35] of several tensors, each one composed of multiple consecutive training input video frames, is manually computed over the entire training set of a large human action recognition dataset. Then, the proposed pretext task consists in learning to map each training input video frame's low-rank approximation to its original version. A lightweight CNN can be pretrained in this manner and then employed in a CNN-LSTM architecture for regular downstream gesture recognition training. A gesture-based HRI dataset was employed for quantitative evaluation purposes. Results indicate that CNN parameter set initialization using the proposed pretext task outperforms, in gesture recognition test set accuracy, both traditional ImageNet initialization and a competing pretext task initialization. The method comes with zero runtime overhead, allowing us to use a very lightweight neural architecture that can be executed in real-time on embedded AI hardware.

II. Related Work

SSL methods can be divided into three categories, based on the data type used for pretext and downstream training: *image-based*, *video-based pretext* and *video-based*. The first two types focus on learning image representations in pretext pretraining and exploiting them on image-related downstream tasks. Their difference is that purely image-based/video-based pretext methods use image/video data in pretext pretraining, respectively. On the other hand, in video-based SSL approaches both the pretext and the downstream task concern videos.

Regarding image-based SSL, [2] adopts one of the most widely studied pretext tasks, jigsaw puzzle solving. The original image is spatially decomposed along a 3×3 grid and transformed into a jigsaw puzzle by shuffling the patches. During pretext pretraining, the DNN concurrently learns to identify the permutation indices and classify the original images. The jigsaw task is adapted in [21], so that the constructed representations of an image and all of its permutations are similar, while at the same time differing from other image representations, thus constituting them transformation-invariant. In [22], the jigsaw task is handled by a siamese CNN that processes the various puzzle patches independently up until its fully connected layers, while in [23] occluding tiles from other random images and knowledge transfer to networks of different architectures and depth are employed.

Image colorization is performed in [42], not aiming to reproduce the original image colors, but instead aiming to produce plausible colorized versions of the input grayscale images. Predicting image rotation and exemplar are the two self-supervised learning methods extended to semi-supervised in [41], while rotation prediction is also handled as a classification problem in [6]. One-shot view grid prediction from a single view is the pretext task posed in [10], enforcing the network to reproduce the 3D shape of an unseen object based solely on one 2D view of itself. In [39] representations are learnt that are capable of discriminating among individual instances, considering that every image belongs to a different class. In [31] synthetic RGB images are leveraged for estimating instance contour, depth and surface normal, while at the same time adapting them to real world data. A different multitask approach is [5], advocating the combination of multiple pretext tasks (relative position, colorization, exemplar, motion segmentation) for inducing richer image representations, while in [43] split-brain autoencoders are proposed that force two sub-networks to perform cross-channel prediction.

Numerous pretext tasks for obtaining good image representations that leverage spatiotemporal information present in videos have been proposed over time, since videos can capture scene dynamics unavailable in static images. In [33] image-, shot- and video-level context is exploited to instill information from the different video granularities to their representations, also employing a small amount of labeled data, while in [20] motion cues are combined with images in order to predict possible changes over time. The task of sequence sorting is adopted in [12], fusing pairwise extracted features with an Order Prediction Network. Foreground-background segmentation is studied in [28], grouping foreground pixels with the aid of optical flow to create ground-truth. Finally, in [38], a siamese-triplet learns to predict similar representations for two tracked image patches of the same video and different representations for randomly sampled patches.

In video-based SSL, where the downstream task also concerns videos, an important milestone was [7], i.e., a pretext task for learning spatiotemporal video embeddings by predicting the future content. In contrast, video pace prediction is employed in [37], based on the assumption that video content understanding is a prerequisite for distinguishing between edited video variants with different pace. In [40], video clip order prediction is employed by pairwise concatenating video clip features extracted using 3D CNNs, while in [11] a *Space-Time Cubic Puzzles* pretext task is presented, i.e., the equivalent of image jigsaw puzzle for videos, only fusing clip features at the final fully-connected network layers and considering it as a permutation problem. This is also adopted by [1], which

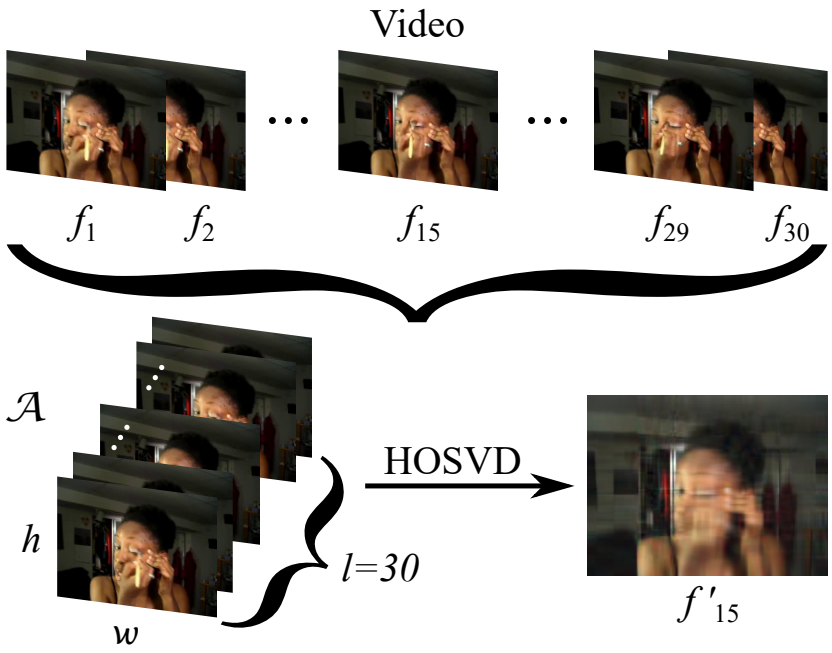

Fig. 1. Training dataset preparation pipeline for the HOSVD compression pretext task.

proposes a novel permutation strategy that preserves spatial coherence. In an entirely different approach, [36] aspires to learn spatiotemporal representations by employing prediction of video sequence motion and appearance statistics.

III. HOSVD Compression Pretext Task

The method proposed in this paper is in essence a video-based SSL approach, since both the pretext and the downstream task act on video data, but the goal is to obtain good per video frame representations, as is typically the case in video-based pretext SSL approaches. Thus, in pretext pretraining, the CNN learns to embed temporal information from video content into the individual video frame descriptions it outputs. Subsequently, this pretrained CNN can be used as a feature extraction backbone in a typical CNN-LSTM setting and trained regularly in an end-to-end manner for a video classification downstream task, such as gesture recognition from RGB camera feed. The proposed *HOSVD compression* pretext task is described below.

Higher Order Singular Value Decomposition (HOSVD) is a well-known generalization of matrix SVD to tensors and constitutes a popular way to solve the Tucker Decomposition [34]. A N-order tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ can be decomposed into a small *core tensor* $\mathcal{D} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N}$, along with a set of orthogonal projection matrices, the so-called *factor matrices* $\mathbf{U}_n \in \mathbb{R}^{d_n \times R_n}$, $n = 1, 2, \ldots, N$, spanning one high-variance subspace for each mode of the core tensor. It holds that:

$$\mathcal{A} \approx \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_N=1}^{R_N} \sigma_{r_1 r_2 \ldots r_N} \mathbf{u}_1^{(r_1)} \otimes \mathbf{u}_2^{(r_2)} \otimes \cdots \otimes \mathbf{u}_N^{(r_N)}, \tag{1}$$

where $\otimes$ is the outer product operator, $\sigma_{i,j \ldots n}$ are the singular values, contained in the core tensor, and $\mathbf{u}_i^{(r_i)}$ are the singular vectors per mode, contained in the corresponding factor matrices [3]. Eq. 1 decomposes tensor $\mathcal{A}$ into a linear combination of rank-one tensors. Thus, if $\mathcal{A}$ is an original data matrix, then $\mathcal{D}$ holds a compact representation of the data, which are then essentially projected onto the given basis factors, in order to form a least-squares approximation of $\mathcal{A}$.

Similarly to the way low-rank approximation can be applied for 2D image compression using matrix SVD [13], [18], [19], a desired subset of core tensor slices and corresponding factor matrix columns can be discarded, separately along each tensor dimension, before approximately reconstructing $\mathcal{A}$ according to Eq. 1. This is implemented by independently setting *tensor ranks* $R_i, i = 1, 2, \ldots, N$ to smaller values, thus extending the concept of low-rank approximation to tensors. Such a process is called *HOSVD truncation* and forms the foundation of the proposed approach.

First, 3D volumes of l consecutive input training video frames are created using a sliding window setting, across the entire pretext training set. Their truncated-HOSVD approximations are subsequently employed as input video frames to the supervised pretext task, with the corresponding original/uncompressed video frames functioning as the respective pseudo-labels. A CNN trained on this task attempts to reconstruct each original video frame, given a spatiotemporally compressed version of itself as input. Thus, it learns to incorporate temporal information from the entire volume of l time instances into each individual video frame representation it computes, by emphasizing visible regions that depict moving objects; their visual details are typically suppressed by spatiotemporal compression. The pretext training dataset

preparation process is described below.

For each training input video V composed of n video frames $f_i,\ i = 1, \ldots, n$, a sliding window of length $l = 30$ and step $s = 15$ video frames is applied in order to estimate the spatiotemporally compressed version of its middle video frame, separately per sliding window. This is performed by calculating the truncated HOSVD of a 3rd-order $w \times h \times l$ tensor $\mathcal{A}$, where w/h are the spatial width/height (in pixels) of the dataset videos, respectively, as shown in Figure 1. Thus, temporal information is embedded to the approximated/compressed middle video frame, by eliminating visual details of visible moving objects.

After empirical investigation of truncating the three tensor modes to different ranks, applying spatial and temporal truncation alone, as well as joint spatiotemporal truncation, we ended up with different SSL pretext input dataset variants which were independently evaluated.

IV. Quantitative Evaluation

As in the vast majority of SSL approaches for videos, we chose "split1" of the UCF101 human action recognition video dataset [32] for pretext pretraining. UCF101 is a benchmark dataset comprised of 13320 videos of 101 different action categories with spatial resolution 320×240 pixels. Transferability of the features produced through pretext pretraining is evaluated by regular supervised training on a gesture recognition task using a randomly selected subset of the recently introduced AUTH UAV Gesture Dataset [29], designed for HRI with autonomous drones/Unmanned Aerial Vehicles (UAVs). The employed subset used consists of 275 videos in total, captured both indoors and outdoors, with static and moving cameras, divided into 6 classes. Its data were amassed from several preexisting datasets, thus spatial resolutions vary.

(a) ImageNet (b) HOSVD compression

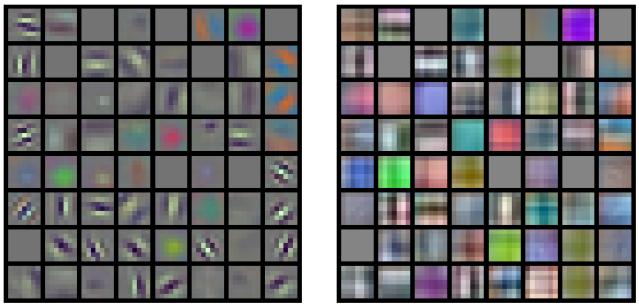

Fig. 2. Conv1 filter visualization.

A. Implementation Details

We employed a ResNet-18-based [8] CNN autoencoder for pretext pretraining. Starting with a regular ResNet-18 architecture in the role of a CNN encoder, which is the desired feature extraction backbone network, a mirror CNN decoder (composed of consecutive deconvolutional layers) was inserted just before the final average pooling layer. Thus, a CNN autoencoder was obtained and trained using binary cross-entropy (BCE) loss function for minimizing the original video frame reconstruction loss. Encoder parameters were initialized with ResNet-18 weights pretrained for whole-image classification on the ImageNet dataset [4].

Training input RGB video frames from "split1" of the UCF101 dataset were normalized using ImageNet dataset mean and std values, resized to 256×256 and afterwards randomly cropped to 224×224. Then, the proposed process for pretext task data preparation (described in Section III) was followed. The batch size used for pretext pretraining was 64 and a SGD optimizer with momentum 0.98 and weight decay of 0.001 was employed. The initial learning rate was 0.1, decaying every 30 epochs, and training was stopped after 200 epochs.

After pretext pretraining, the ResNet-18 encoder was attached as backbone feature extraction network to an LSTM [9] network with input dimension equal to 4096, i.e., the length of the feature vector produced by the backbone, 2 layers of 128 neurons each, and unrolling for 15 time steps. The entire CNN-LSTM architecture was trained end-to-end for the desired gesture recognition task on the AUTH UAV Gesture Dataset, using truncated backpropagation through time (BPTT) for 100 epochs. Learning rate was set to 0.001, decaying every 30 epochs. The batch size used was 20 and an SGD optimizer with momentum 0.98 and weight decay of 0.001 was employed.

Before downstream training, all gesture input videos were first set to a temporal length equal to 15 video frames, so that all samples are equally handled by the LSTM network. This was done by random video frame subsampling, for videos of length larger than 15 video frames, and random video frame duplication for videos of less than 15 frames.

TABLE I
Quantitative evaluation results on the employed subset of the AUTH UAV Gesture Dataset, using the Correct Classification Rate (CCR) metric.

Initialization	CCR
Random	56.90 %
ImageNet	60.19 %
RotNet [6]	60.00 %
HOSVD (proposed)	**64.08 %**

B. Evaluation Results

The proposed HOSVD compression pretext task was examined with regard both to the nature of the representations it produces and their transferability to a gesture recognition downstream task. A widespread approach to qualitatively demonstrate the image representations produced by CNNs pretrained on pretext tasks is by visualizing the filters of their first convolutional layer ("Conv1"). Thus, Figure IV compares the Conv1 filters obtained by: a) pretraining the ResNet-18 encoder alone on ImageNet for whole-image classification, and b) attaching to the ResNet-18 encoder a corresponding decoder and pretraining the overall model for the HOSVD compression

pretext task. As expected, ImageNet classification pretraining provides good gradient/edge detection convolutional filters, while UCF101 HOSVD compression pretraining provides filters sensitive to spatial texture frequency and appearance. Thus, initializing the CNN for the downstream task using the proposed method renders the model able to better distinguish between visible human body regions (which are characterized by specific texture/appearance patterns) from the background, or from very differently looking objects.

Quantitative evaluation results on the selected subset of AUTH UAV Gesture Dataset are shown in Table I. Evidently, initializing the ResNet-18 part of the CNN-LSTM architecture for the desired gesture recognition downstream task using the parameter set obtained by HOSVD compression pretraining, gives rise to higher Correct Classification Rate (CCR) than both ImageNet initialization and than using competing pretext pretraining [6]. The reported HOSVD compression pretraining results were obtained by retaining 10%/70% of the spatial/temporal dimensions, respectively, when applying HOSVD truncation during pretext dataset preparation. These values were selected after a careful ablation study, which showed that further amplification of spatial/temporal distortion in pretext input training data led to deteriorating downstream test classification rates, possibly due to the excessive loss of image details.

Overall, the proposed pretraining method significantly increased test set CCR for the downstream gesture recognition task, without any modifications to the neural architecture or the input data. The achieved accuracy still lags compared to approaches that compute and classify sequences of 2D human body skeletons [27], instead of operating directly on the RGB images, but such approaches require more complex and slower DNNs. In our case, since the employed CNN-LSTM architecture was very lightweight, the overall system was able on average to process per second approximately 52 sequences, each one composed of 15 video frames, and thus output an equal number of gesture predictions per second. Evaluation was performed on an embedded AI compute board (nVidia Jetson Xavier), which is suitable for robotic platforms. Thus, employing the proposed pretext task permits us to rely on less complex and much faster neural models, capable of being executed on-board an autonomous system in real-time.

V. Conclusions

This paper presented a novel pretext task for pretraining a CNN under a video-based self-supervised learning (SSL) setting, where the intention is to later employ this model as a per video frame feature extraction backbone for a gesture recognition downstream task, in the context of a CNN-LSTM architecture. The proposed HOSVD-based spatiotemporal compression pretext pretraining method, relying on tensor decomposition and low-rank approximation, was shown to provide better CNN parameter set initialization than typical pretraining for whole-image classification on ImageNet, as well as than a competing SSL approach. Gesture classification accuracy was significantly augmented in a relevant gesture recognition dataset for Unmanned Aerial Vehicle (UAV) handling, without any modifications to the lightweight neural architecture or the input data, while the overall system was shown to perform in real-time on embedded AI hardware. Thus, the proposed method seems to be appropriate for Human-Robot Interaction applications, where gesture recognition needs to be performed on-board an autonomous system.

References

[1] U. Ahsan, R. Madhok, and I. Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

[2] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] J. Chen and Y. Saad. On the tensor SVD and the optimal low rank orthogonal approximation of tensors. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1709–1734, 2009.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[7] T. Han, W. Xie, and A. Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV)*, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[10] D. Jayaraman, R. Gao, and K. Grauman. ShapeCodes: self-supervised feature learning by lifting views to viewgrids. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[11] D. Kim, D. Cho, and I. S. Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[12] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICIP)*, 2017.

[13] X. Ma, X. Xie, K.-M. Lam, J. Hu, and Y. Zhong. Saliency detection based on Singular Value Decomposition. *Journal of Visual Communication and Image Representation*, 32:95–106, 2015.

[14] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas. High-level multiple-UAV cinematography tools for covering outdoor events. *IEEE Transactions on Broadcasting*, 65(3):627–635, 2019.

[15] I. Mademlis, V. Mygdalis, N. Nikolaidis, and I. Pitas. Challenges in autonomous UAV cinematography: An overview. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[16] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina. Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments [applications corner]. *IEEE Signal Processing Magazine*, 36(1):147–153, 2018.

[17] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina. Autonomous UAV cinematography: a tutorial and a formalized shot-type taxonomy. *ACM Computing Surveys (CSUR)*, 52(5):1–33, 2019.

[18] I. Mademlis, A. Tefas, and I. Pitas. Regularized SVD-based video frame saliency for unsupervised activity video summarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] I. Mademlis, A. Tefas, and I. Pitas. A salient dictionary learning framework for activity video summarization via key-frame extraction. *Information Sciences*, 432:319–331, 2018.

[20] A. Mahendran, J. Thewlis, and A. Vedaldi. Cross pixel optical-flow similarity for self-supervised learning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2018.

[21] I. Misra and L. Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[23] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] S. Papadopoulos, I. Mademlis, and I. Pitas. Neural vision-based semantic 3D world modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[25] S. Papadopoulos, I. Mademlis, and I. Pitas. Semantic image segmentation guided by scene geometry. In *IEEE International Conference on Autonomous Systems (ICAS) (accepted for presentation)*, 2021.

[26] C. Papaioannidis, I. Mademlis, and I. Pitas. Fast single-person 2D human pose estimation using multi-task convolutional neural networks. In *(submitted)*, 2021.

[27] C. Papaioannidis, D. Makrygiannis, I. Mademlis, and I. Pitas. Learning fast and robust gesture recognition. In *EURASIP European Signal Processing Conference (EUSIPCO) (accepted for presentation)*, 2021.

[28] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] F. Patrona, I. Mademlis, and I. Pitas. An overview of hand gesture languages for autonomous UAV handling. In *(submitted)*, 2021.

[30] F. Patrona, I. Mademlis, A. Tefas, and I. Pitas. Computational UAV cinematography for intelligent shooting based on semantic visual analysis. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019.

[31] Z. Ren and Y. Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[33] M. Tschannen, J. Djolonga, M. Ritter, A. Mahendran, N. Houlsby, S. Gelly, and M. Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[35] N. Vannieuwenhoven, R. Vandebril, and K. Meerbergen. A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing*, 34(2):A1027–A1052, 2012.

[36] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[37] J. Wang, J. Jiao, and Y.-H. Liu. Self-supervised video representation learning by pace prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[38] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[39] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[42] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[43] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.